DOCUMENT RESUME

ED 334 218 TM 016 676

AUTHOR Terwilliger, James S.
TITLE Distinctions between Item Format and Objectivity in Scoring.
PUB DATE Apr 91
NOTE 24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classroom Techniques; *Educational Testing; Higher Education; *Measurement Techniques; Objective Tests; Questioning Techniques; *Scoring; Teacher Education; Test Construction; Test Format; Testing Problems; *Test Items; Textbook Content
IDENTIFIERS Free Response Test Items; *Paper and Pencil Tests

ABSTRACT

This paper clarifies important distinctions in item writing and item scoring and considers the implications of these distinctions for developing guidelines related to test construction for training teachers. The terminology used to describe and classify paper and pencil test questions frequently confuses two distinct features of questions: response format, and objectivity in scoring. There are three major response formats: (1) choice response; (2) free response; and (3) any combination of choice response and free response. Objectivity may be operationally defined by asking subject matter experts to indicate which responses should be accepted as correct. The consensus among the experts provides an empirical definition of the objectivity of the item. Results of a review of 10 commonly used textbooks on classroom measurement are discussed, which reveal inconsistency and confusion in the treatment of issues related to objectivity. Textbooks often imply that objectivity is guaranteed by specific response formats; illustrations are provided to show this is not true. Two figures and one table illustrate the discussion. A bibliography of the 10 textbooks on classroom measurement that were reviewed is appended. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334218

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JAMES S. TERWILLIGER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Distinctions Between Item Format and Objectivity in Scoring

James S. Terwilliger

University of Minnesota

Paper presented at annual meetings of National Council on Measurement in Education, Chicago, April 5, 1991.

TM016676

BEST COPY AVAILABLE

Distinctions Between Item Format and Objectivity In Scoring

James S. Terwilliger, University of Minnesota

#2,#3

ABSTRACT

The terminology employed to describe and classify paper and pencil test questions frequently confuses two distinct features of questions: (a) response format and (b) objectivity in scoring. There are three major response formats: 1) choice response, 2) free response and 3) some combination of choice and free responses. Objectivity may be operationally defined by asking subject matter experts to indicate which response(s) should be accepted as correct. The consensus among the experts provides an empirical definition of the objectivity of the item. Textbooks often employ that objectivity is guaranteed by specific response formats. Illustrations are provided to show this is not true.

Introduction

Textbooks and references commonly employed in courses on classroom assessment typically contain several chapters on the planning and construction of paper and pencil tests. As a means of organizing the discussion of the diverse item types which are most frequently employed in written tests, textbook authors use a variety of terms, e.g. "objective type" item, essay question, short answer item, completion item, choice response item, supply type question, etc. Unfortunately, there is no consensus among authors in the use of terminology for describing test items. This is often a source of confusion to students.

A more serious problem is that authors often employ systems for classifying test questions which confuse differences among questions in terms of the format in which they are presented and presumed differences among questions in terms of inter-judge agreement in evaluating responses to the questions. This tends to obscure issues associated with item writing with issues associated with item scoring. The purpose of this paper is to clarify important distinctions in item writing and item scoring. Further, the implications of these distinctions for developing guidelines related to test construction will be noted.

Item Response Formats

The most obvious way to describe variations among items in written tests is in terms of the nature of the response which is called for by the item. Textbook authors describe these differences using terms such as choice response items (including

true-false, multiple-choice, matching, etc.), short answer/completion questions, and essay questions (both restricted and extended response varieties). An implicit assumption in most textbooks is that item response formats are either choice response (select type) or free response (supply type).

The possibility of developing items which require both a choice and a free response is rarely considered despite the fact that such items may be perfectly appropriate. For example, asking students to determine whether a statement is true or false and, if false, indicating why that is the case. Or asking students to determine whether the punctuation, grammar or spelling in a specific written passage is correct or incorrect and, if incorrect, making proper corrections or revisions.

Figure 1 presents a general classification scheme for describing test items according to various response formats. An effort has been made to list specific item formats in increasing order of complexity within both the choice response and free response categories.

Insert Figure 1 about here

Objectivity and Item Scoring

Definitions of Objectivity

To gain an understanding of the concept "objectivity" it is useful to consider first how the notion is described by philosophers. Newell (1986) notes that there are two distinct

"faces" of objectivity. The first face he associates with the philosophy of Kant.

> It gives objectivity two roles. First, an ontological role ascribing objectivity to 'objects', or particular bodies, entities, complexes or states of affairs existing apart from perceptions more or less continuously in space and time, falling under the heading of 'objective particulars'. And second, an epistemological role in which objectivity is ascribed to items of a different sort, beliefs, judgments, propositions or products of thought about what is really the case, forming a general class of 'objective judgments'. The two roles become linked when objective particulars are seen as the topics of objective judgments and, thus linked, 'objectivity' becomes an explanatory notion. (p.16)

This notion of objectivity has been called ontological objectivity or veridicality by Eisner (1991) since it deals with "seeing things the way they are".

The second face of objectivity is less technical and addresses judgmental processes. It is described by Newell (1986) as follows:

> Objective judgments are contrasted with prejudiced, biased or dogmatic judgments and objectivity is associated with impartiality, detachment, disinterestedness and a willingness to submit to standards of evidence. It is distinctive of this view that objectivity attaches to persons through their actions. What makes a judgment

> objective is not something special about outer objects, but something special about people's practices. Seen in this way objectivity is an inter-personal notion, giving sense to the idea that a person may be, or may become objective by the exercise of a disposition to act within the constraints of a wider social practice. (p. 17-18)

Eisner (1991) refers to this notion as procedural objectivity since it is premised upon the "use of a method that eliminates, or aspires to eliminate, the scope of personal judgment in the description and appraisal of a state of affairs".

It is interesting to note that both Newell (a philosopher) and Eisner (an educational researcher), although writing from quite different perspectives, regard the first conception of objectivity as of limited practical use. As Newell (1986) states:

> To explain 'things as they really are' by calling them 'objective' accomplishes nothing without a prior grasp of the difference between 'real things' and representations, and if we already have a sense of that distinction then there is not need to invoke the notion of objectivity to provide it. Nor does it help to define a subjective-objective difference by an inner-outer contrast. Defined in this way, the question "What is the difference between the subjective and the objective?' is answered comprehensively by answering the counterpart question 'What is the difference between the inner and the outer?' Every problem

> or query about the one is a problem or query about the other, and nothing of an explanatory nature is gained by the translation. Objectivity's first face makes objectivity a redundant concept. (p. 19-20)

With regard to procedural objectivity, Eisner (1991) observes:

> One of the commonest examples of such a method is the objectively scored achievement test. Once the test has been constructed, identifying a correct or incorrect response does not require interpretation. Although there are interpretive issues at stake at the level of test construction, at the level of scoring, the optical scanner will do. Since scoring requires no judgment, it is procedurally objective -- hence, we say that we have an objective test or an objective method for scoring responses. (p. 44)

Later he provides another example.

> Procedural objectivity is also secured by training judges to make identical judgments about the objects of their judgment. This is done by specifying the criteria that are to be employed and by assigning numbers to qualities along a continuum displaying excellence on each criterion. When judges are successfully trained in the application of the criteria, correlations among the judges' scores will be high; ideally, they will be uniform. In such circumstances, even when measurement is not used in a literal sense, we say

> that the judges have used objective or nearly objective scoring procedures. We believe that judges at the Olympic Games are not making merely subjective judgments about the quality of the dives they score. As differences among judges' scores increase, our confidence in the judges' judgment decreases. Consensus breeds confidence in the objectivity of the judgments rendered. (p. 44-45)

It is important to note that the first example given by Eisner needs modification. The notion that "identifying a correct or incorrect response does not require interpretation" is never true. All test items are presented using symbols (words, numbers, graphs, etc.) which require interpretation both on the part of the test taker and on the part of the item writer. Frequently, there are also a variety of contextual factors (e.g. implicit assumptions made by the item writer) which add to the cognitive complexity presented by the item. The interpretation of symbols in a test item and the associated contextual factors have a direct bearing upon the determination of whether a response is judged correct or incorrect. Therefore, it is a gross oversimplification to imply that objectivity is equivalent to using an optical scanner for scoring responses.

The second illustration given by Eisner is more informative because it addresses objectivity in a more fundamental way. The notion that a set of judges should produce a high degree of agreement based upon: a) extensive training and b) shared criteria for judging complex performances is a useful model for

considering objectivity in educational measurement. First, there is explicit recognition that those making judgments are "expert" by virtue of their training and familiarity with standards of excellence. Second, the idea of consensus among judges acknowledges that objectivity is subject to empirical verification. Third, since agreement or consensus is a matter of degree, objectivity should not be regarded as an "all or none" concept. Rather it should be recognized that responses to test items can be scored with varying degrees of objectivity.

Objectivity and Classroom Testing

Variations on phrases such as "objective test" and "objective type items" are commonly employed in references on educational measurement. Yet discussions of the meaning of the concept "objectivity" and its relevance to sound measurement practices are very limited in texts on measurement. The index of the most recent edition of Educational Measurement (Linn, 1990) contains only a single reference to the term "objectivity". That occurs in the context of a brief discussion of philosophical issues associated with test validity. The index does not contain the terms "objective item" or "objective test".

In contrast, the index to an earlier edition of the same reference volume (Thorndike, 1971) lists eight references to "objective test" but no reference to the term "objectivity". In fact, Chapter Three is entitled "Planning the Objective Test" but there is no definition of the phrase "objective test" in that chapter. In Chapter Five on "Writing the Test Item", the

following definitions are given:

> An item may be defined as a scoring unit. An exercise may be defined as a collection of items that are related structurally. For example, a matching exercise may consist of five items; a reading passage and the number of items based upon it constitute a test exercise. An objective item or exercise is one that can be scored by mechanical devices or by clerks who have no special competence in the field. (Wesman, p. 81)

This definition of an objective item is flawed in the same way as the earlier example given by Eisner (1991) - it confuses the use of a mechanical scoring procedure with an empirical demonstration that there is a high level of consensus among experts with respect to the correct answer.

A review of commonly used texts on classroom measurement reveals the same inconsistency and confusion in the treatment of issues related to objectivity. Inspection of a sample of ten such texts[1] readily available to the author reveals the results shown in Table 1. Entries in the cells and along the margins of Table 1 are frequencies. As can be seen in the table, three of the texts do not define "objective item" in terms of consensus. One gives no definition and the other two provide definitions of

---

Insert Table 1 about here

---

[1] The ten texts reviewed are listed in Appendix A.

"objective item" (or "objective test") in terms of the specification of mechanical or clerical scoring rather than in terms of agreement among experts.

Eight of the ten texts employ the dichotomy "objective vs. essay" when discussing various types of paper and pencil questions. This dichotomy apparently represents the "conventional wisdom" among authors of texts on classroom assessment. This dichotomy needs closer scrutiny.

Objective Versus Essay Questions

In the eight texts which employ the distinction between objective and essay questions, the dichotomy first appears in a chapter on planning classroom tests. One or more subsequent chapters are devoted to "objective tests" (true-false, multiple-choice, completion, etc.) and a single chapter is devoted to essay questions. As noted in Figure 1, essay refers to the format for responding, not the nature of the scoring. Although there is ample evidence that different readers frequently show only modest agreement when scoring essay responses, this lack of objectivity should not be viewed as inherent to the essay format.

Coffman (1971) reviewed data based upon a variety of studies which indicate that it is possible to achieve respectable inter-rater agreement in the scoring of essay questions. This was accomplished through a combination of approaches such as developing clearly defined rating scales, specifying explicit scoring criteria, using model answers as a basis for scoring, and, most importantly, providing extensive training for readers.

In other words, given sufficient resources, essay questions can be scored with a reasonable degree of objectivity. The problem, of course, is that such resources are rarely available.

Even if it is conceded that, as typically employed, essay questions are scored in a highly subjective fashion, it does not follow that all other response formats result in highly objective scoring. The scoring key developed by one subject matter expert may differ from the key developed by another expert. This can happen regardless of the response format.

The likelihood of disagreements in scoring is greater when questions are designed to tap higher order thinking skills than when items measure only knowledge and/or basic skills. One reason why item formats like true-false, multiple-choice and completion are automatically considered as "objective" is that they are most commonly used to measure factual knowledge. As most textbook authors note, a heavy emphasis upon basic knowledge outcomes is neither necessary nor desirable. Teachers should be encouraged to construct items that measure cognitively complex outcomes. However, in so doing, they should also be encouraged to check their scoring keys against judgments of others.

Regardless of the response format employed, a fundamental issue in judging the quality of a test item or question is the degree of agreement among subject matter experts with respect to the acceptability or correctness of a given response. Where there is high consensus among authorities in judging the quality of a response to a question, a scoring system which is objective

can be designed. On the other hand, where there is little consensus among experts concerning the appropriateness of a response the scoring of the item will be subjective.

Obviously, consensus is a matter of degree ranging from extremely low (e.g. less than 50 percent of the experts agree on the quality of a specific response) to extremely high (e.g. over 90 percent of the experts agree on the correctness of a specific response). Consequently, the objectivity a test item is also a matter of degree rather than "all or none".

A commonly held misconception is that the designation of one response as the "correct" response by the item writer automatically qualifies the item as "objective". It is further assumed that the existence of a scoring key which specifies the acceptable answers to all items in a test makes the test "objective" because the same key can be applied consistently to the responses of each student. This view is reflected in the following comments by Wiersma and Jurs (1985).

> The scoring of multiple-choice and true-false items is reliable because two independent scorers or one person scoring the same test twice should arrive at identical scores. Scorers are not called upon to make judgments; they only have to determine whether the answers on a test or answer sheet match the preferred responses given on the scoring key. Since no judgments are needed, the scoring can be done by clerks or machines, rather than teachers, counselors, or other highly trained staff.

> Note the distinctions among objectivity, reliable scoring, and test reliability. Scoring is objective because it requires no judgments or inferences from the scorer; thus, it is very likely that the item scoring will be consistent. This means that the scoring is also reliable. (p. 124)

The construction of a key by the test designer does not make the test objective. To support the claim that a test is objective it is necessary to provide data regarding consensus among subject matter experts with respect to the correct answers on the test. This suggests that a teacher should ask two or three colleagues in the same subject matter area to respond to the test items which have been prepared. Those items on which there is high consensus in responses are, by definition, objective items. Items on which there is low consensus should be reviewed in an effort to determine the source of the disagreement.

Many textbook authors use the expression "objective-type" item but few note that the objectivity of an item is subject to empirical test and is not guaranteed by virtue of the fact that the item uses a short answer or choice response format. Authors also frequently imply that "objective" items have only one acceptable answer. This is clearly not the case if it can be demonstrated that experts agree that two or more responses are acceptable for a given completion or short answer question.

Figure 2 shows a framework for describing items with respect to format and scoring. Illustrative items representing various

combinations of response formats and scoring consensus are also presented. This figure makes clear that the response format and the nature of the scoring of an item are separate considerations. This is acknowledged by Ebel and Frisbie (1991) who state:

> Whichever type examiners decide to use, they should strive to make their measurements as objective as possible. A measurement is objective to the extent that it can be verified independently by other competent measurers. It is conceivable that measurements obtained from a good essay test could be more objective in this sense than measurements obtained from a poor multiple-choice test. (p. 116)

Insert Figure 2 about here

Implications for Training Teachers

Distinctions between item formats and item scoring procedures are extremely important in clarifying practical measurement issues that teachers need to understand. Textbooks on classroom assessment frequently draw contrasts between "essay vs. objective" type items. As can be seen in Figure 2, this is a confusing dichotomy. Reasonably objective scoring techniques can be developed for essay questions through the use of model papers and analytic scoring keys. On the other hand, choice response and short answer questions frequently fail to produce responses on which subject matter experts agree. Such questions should not be labeled "objective".

The importance of using items which can be scored objectively cannot be overemphasized. Regardless of the format teachers employ in designing classroom tests, the issue of expert agreement in judging the responses is fundamental. This point is made forcefully by the authors of two of the texts listed in Appendix A. Ebel and Frisbie (1991) state:

> In considering the relative merits of essay, problem, and objective tests, it is important to remember that the only useful component of any test score is the objectively verifiable component of it. To the degree that a test score reflects the private, subjective, unverifiable impressions and values of one particular scorer, the score is deficient in meaning, and hence in usefulness, to the student who received it or to anyone else who is interested in using it. (p. 116)

Hopkins, Stanley, and Hopkins (1990) reflect the same viewpoint in the following comment in the introductory pages of their text.

> It is desirable to increase the objectivity of testing, interviewing, rating, and similar enterprises that often are pursued quite subjectively.
>
> Increasing the objectivity of the assessment and evaluation of human behavior is a major goal of this book. We are interested in increasing objectivity not for its own sake but because, as later chapters will make clear, the validity of a measure or evaluation is usually enhanced as it becomes less subjective. (p. 8)

Teachers should be encouraged to submit their test items and questions to their colleagues for critique and comment as a standard practice. Alternatively, teachers might wish to undertake test construction as a cooperative project. An advance discussion of problems likely to be faced in scoring tests is a good way to assure both greater objectivity in scoring and higher quality test items.

There is a final important practical implication of objectivity for classroom teachers. To the extent that objectivity is a goal in testing (there appears to be near unanimous agreement among text authors that it should be), a great deal of additional effort will be needed by those who employ essay questions to assure that the goal is attained. Teachers will need to take great care in structuring essay questions, in designing model papers and scoring guides, and in eliminating extraneous factors (handwriting, punctuation, grammar, etc.) which are not regarded as outcomes to be assessed. It is clear that if such efforts are made, the time invested in the construction and scoring of essay questions would increase dramatically. Consequently, the rules of thumb which appear in many textbooks concerning the relative time advantages in preparing and scoring essay questions would no longer apply. Instead, the point would be made that the production of objectively scorable tests requires a substantial commitment of time and energy, regardless of the response format employed.

References

Coffman, W. (1971). "Essay Examinations". In Thorndike, R. (Ed.) Educational Measurement (2nd ed.) Washington, D.C.: American Council on Education. p. 271-302.

Ebel, R. and Frisbie, D. (1991). Essentials of Educational Measurement. (5th ed.) Englewood Cliffs, N.J.: Prentice Hall.

Eisner, E. (1991). The Enlightened Eye. New York: Macmillan.

Hopkins, K., Stanley, J., and Hopkins, B. (1990). Educational and Psychological Assessment and Evaluation. (7th ed.) Englewood Cliffs, N.J.: Prentice Hall.

Linn, R. (Ed.) (1990). Educational Measurement (3rd ed.) New York: American Council on Education - Macmillan.

Newell, R. (1986). Objectivity, Empiricism and Truth. London: Routledge and Kegan Paul.

Thorndike, R. (Ed.) (1971). Educational Measurement (2nd ed.) Washington, D.C.: American Council on Education.

Wiersma, W. and Jurs, S. (1985). Educational Measurement and Testing. Boston: Allyn and Bacon.

Figure 1
Classification of Items by
Response Format

I. Free Response Format (Supply Type)

A. Short answer/completion item where response is a:
1. single symbol or word.
2. brief phrase (up to five words).
3. single sentence.

B. Problem variety item involving a:
1. single-step solution.
2. multiple-step solution.

C. Restricted response essay question where response is limited to:
1. a single paragraph.
2. two or three paragraphs (about one-half page).

D. Extended response essay question where response consists of written discourse on a major topic that is several pages in length.

II. Choice Response Format (Select Type)

A. Two-choice items
1. True-False
2. Right-Wrong
3. Yes-No
4. Etc.

B. Three-choice items
1. Any two-choice format in combination with "Impossible to say."
2. Increase, Same, Decrease
3. Etc.

C. Four- and five-choice items (e.g. multiple choice format)

D. Matching items

E. Rearrangement items

III. Combination free and Choice Response Formats

A. Two choices in combination with free response (e.g. interlinear exercises)

B. Three choices in combination with free response (e.g. increase, same, decrease and give reason)

C. Other combinations (e.g. multiple choice with explanations of choice made)

Figure 2

Classification of Items by Format and Scoring

SCORING	RESPONSE FORMAT: Choice Response	Free Response	Combination Choice and Free Response
Objective (Hi Consensus)	A test which yields results which are highly consistent is said to be a. valid. *b. reliable. c. accurate. d. dependable.	What term is used to describe a test which yields consistent results? Acceptable responses: a. Reliable b. Stable c. Internal Consistency	Part 1: Choice Response T *F A test which yields highly consistent results is a valid test. Part 2: Free Response To make the item true "valid" should be replaced with "reliable".
Subjective (Lo Consensus)	Which of the following is the best method for improving the reliability of a test? a. increasing the size of the sample taking the test b. increasing the number of items in the test c. improving the clarity and precision of wording of the test items d. increasing the time allowed to take the test Note: Depending upon circumstances, b, c, or d could be correct.	How does an increase in the reliability of the test affect the validity of the test? The answer depends upon what type of reliability and validity are considered.	Part 1: Choice Response T F High test reliability is necessary for the test to be valid. Part 2: Free Response The answer depends upon further specification of the type of reliability and validity at issue. For example, high internal consistency is not necessary for a test to have good predictive validity.

NOTE: The correct response is indicated with an *.

Table 1

Results of Analysis of
Ten Texts on Classroom Measurement

		Employs Dichotomy Objective Test vs. Essay Test		
		No	Yes	
Defines "Objective Item" in Terms of Consensus Among Experts	Yes	1	6	7
	No	1	2	3
		2	8	10

Appendix A

Textbooks on Classroom Measurement Reviewed

Carey, L. (1988). Measuring and Evaluating School Learning. Boston: Allyn and Bacon.

Ebel, R. and Frisbie, D. (1991). Essentials of Educational Measurement. (5th ed.) Englewood Cliffs, N.J.: Prentice Hall.

Gronlund, N. and Linn, R. (1990). Measurement and Evaluation in Teaching. (6th ed.) New York: Macmillan.

Hopkins, K., Stanley, J. and Hopkins, B. (1990). Educational and Psychological Assessment and Evaluation. (7th ed.) Englewood Cliffs, N.J.: Prentice Hall.

Kubiszyn, T. and Borich, G. (1990). Educational Testing and Measurement. (3rd ed.) Glenview, IL: Scott, Foresman.

Mehrens, W. and Lehmann, I. (1984). Measurement and Evaluation in Education and Psychology. (3rd ed.) New York: Holt, Rinehard, & Winston.

Nitko, A. (1983). Educational Tests and Measurement: An Introduction. San Diego: Harcourt, Brace Jovanovich.

Noll, V., Scannell, D. and Craig, R. (1979). Introduction to Educational Measurement. (4th ed.) Boston: Houghton Mifflin.

Popham, J. (1990). Modern Educational Measurement. (2nd ed.) Englewood Cliffs, N.J.: Prentice Hall.

Wiersma, W. and Jurs, S. (1985). Educational Measurement and Testing. Boston: Allyn and Bacon.